DOCUMENT RESUME

ED 188 890 SE 031 135

AUTHOR Tobin, Kenneth G.; Capie, William
TITLE Applications of Generalizability Theory to Classroom Process Measures.
PUB DATE Apr 80
NOTE 32p.: Paper presented at the annual meeting of the National Association for Research in Science Teaching (53rd, Boston, MA, April 11-13, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Classroom Research; Educational Research; *Observation; Reliability; *Research Methodology; Science Education; *Statistical Analysis; Teacher Behavior
IDENTIFIERS Generalizability Theory; Science Education Research

ABSTRACT

This paper describes generalizability theory as an extension to the analysis of variance approach to reliability which allows the researcher to attribute variation to multiple sources, and a number of reliability coefficients to be calculated for most observational contexts. The applications of generalizability coefficients to characterize the dependability of data gathered on classroom process variables are pursued. Models are provided for crossed and nested designs for situations involving two and three sources of variation. Several examples are discussed involving data collected with the Teacher Performance Assessment Instruments (TPAI). Other examples discuss applications of generalizability theory to measures of cognitive level of teacher questioning and measures of student engagement. All examples were selected to highlight applications of generalizability theory, which these authors feel is particularly applicable to educational measures which are often affected by a number of identifiable sources of variance. (CS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kenneth G. Tobin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED188890

APPLICATIONS OF GENERALIZABILITY THEORY TO CLASSROOM PROCESS MEASURES

Kenneth G. Tobin
William Capie

Department of Science Education
University of Georgia
Athens, Georgia 30602

Presented at the
National Association of
Research in Science Teaching

Boston, Mass.

April, 1980

SE 031 135

APPLICATIONS OF GENERALIZABILITY THEORY TO CLASSROOM PROCESS MEASURES

INTRODUCTION

The use of direct observation to obtain data on teachers and students is common in educational research and evaluation contexts. Those responsible for data collection share a concern for dependable measures so that confidence can be placed in decisions made on the basis of the data obtained. Latent recognition is given to the existence of multiple sources of variation through the use of mean ratings rather than measures based on a single observer or a single observation period. Research reports suggest that decisions related to the frequency and duration of observation periods, the number of raters to be used and the types of lessons to be observed are usually guided by experience rather than empirical analysis. Reliability coefficients are rarely used to aid decision making and are usually reported post-hoc as correlation coefficients or intraclass coefficients obtained from one-way analysis of variance techniques (e.g., Hoyt, 1941; Ebel, 1951; Cronbach, 1951). Coefficients that have accounted for the multiple sources of variance extant in most observational measures are rarely reported.

Generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) provides a formulation that is applicable to observational measures. As an extension of the analysis of variance approach to reliability, generalizability theory attributes variation to multiple sources, allowing a number of reliability coefficients to be calculated for most observational contexts. The use of generalizability coefficients to characterize the

dependability of data gathered on classroon process variables is an obvious area of application. Data of this type are used in research and evaluation contexts to answer a diverse set of questions. As a consequence, estimates should be obtained on the reliability of measures for answering questions of concern. If possible, a study should be designed to obtain generalizability coefficients that reflect the sources of variation in the data obtained. Studies of this type have been called G-studies to indicate the purpose of obtaining generalizability data (Cronbach et al, 1972). On the basis of coefficients obtained, the data collection procedures can be modified to obtain dependable measures in subsequent D-studies where decisions are made based on the data. Where a G-study cannot be designed, coefficients can be estimated from data already collected in an investigation, to provide a confidence interval on the measures obtained.

In 1972, McGaw, Wardrop and Bunda applied generalizability theory to classroom rating situations with particular emphasis on applications in G-studies. Their paper clearly demonstrated that a number of generalizability coefficients can be calculated to assist in distinguishing random from lawful sources of variation in observational data. Cardinet, Tourneur and Allal (1976) contributed to the application of generalizability theory by emphasizing the symmetry of the model. In a three facet crossed design they illustrated that any of the facets can be regarded as facets of differentiation or facets of generalization. Their discussion of sixteen interpretable generalizability coefficients highlighted the contextual dependence of the coefficients. A set of measures might be quite dependable for certain questions but not dependable for others.

In this paper the symmetrical aspects of generalizability theory, articulated by Cardinet et al, are applied to a range of contexts involving measures on classroom process variables. Models are provided for crossed and nested designs for situations involving two and three sources of variation. Several examples discussed in the paper involve data collected with the Teacher Performance Assessment Instruments (TPAI) (Capie, Johnson, Anderson, Ellett and Okey, 1979). Other examples discuss applications of generalizability theory to measures of cognitive level of teacher questioning and measures of student engagement. Examples were selected to highlight applications of generalizability theory and the coefficients reported may differ from those estimated from more representative data.

BASIC TERMINOLOGY

The sources of variation in a generalizability analysis are referred to as facets, which are analogous to factors in analysis of variance. A variance component is estimated for each facet and a generalizability coefficient is calculated using the formula shown below. Cronbach et al (1972) used the term rho-squared (ρ^2) to designate the generalizability coefficient.

$$\rho^2 = \frac{\text{estimated universe score variance}}{\text{estimated observed score variance}}$$

Rho-squared provides a measure of the dependability of the measures for differentiating among levels of specified facets and generalizing over the levels of others. Variance components are used to estimate both the universe score variance and the observed score variance for the facets of interest.

The facets to be differentiated should always be distinguished from facets of generalization. In many investigations of classrooms, teachers represent a facet to be differentiated. For example, to answer the research question of the relationship of variation in teacher performance to variation in pupil achievement, teacher performance must be differentiated. Thus, a generalizability coefficient should be calculated to ascertain whether the measures are sufficiently reliable to differentiate teacher performance. Raters are also often regarded as a facet to be differentiated. In this case, the magnitude of the variance between raters is compared with variance due to other sources to obtain a generalizability coefficient.

A facet of generalization is a source of variation which affects the measures obtained in an investigation. It includes random errors of measurement and a systematic source of variance. Common examples of facets of generalization include raters, occasions of measurement, etc.

Consider the example where a generalizability coefficient is to differentiate teachers and generalize over raters, categories and occasions.

TWO FACET DESIGNS

The use of generalizability theory with a two facet design will lead to the same results as those obtained for coefficient α (Cronbach, 1951) or intraclass coefficients obtained using Hoyt's procedures. Whilst the Hoyt and Cronbach procedures are not applicable to situations that involve more than two facets, generalizability theory can be applied to contexts involving multiple sources of variation. In this paper the generalizability formulation is initially introduced in a two facet design and is later extended to incorporate three sources of variation.

Crossed Design

In an investigation designed to estimate the generalizability of teacher performance measures, six raters each assessed four teachers with the TPAI. The measure of teacher performance was obtained by summing competency scores related to interpersonal skills and classroom procedures. This design can be represented as fully crossed, with raters and teachers as facets in a two-way analysis of variance. The formulae for estimating the components of variance for this design are contained in Table 1.

(Include Table 1 About Here)

The generalizability coefficient to differentiate teachers is calculated from the formula below. The component of variance attributable to error is divided by k, the number of raters, since the measures of concern are mean performance ratings based on k raters.

$$\rho_t^2 = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \frac{1}{k}(\hat{\sigma}_{res}^2)}$$

The ANOVA table that is applicable to the example is contained in Table 2. Using the variance component estimates from Table 2, a coefficient of .99 was obtained. Differentiation of teachers was quite satisfactory using this procedure.

(Include Table 2 About Here)

A similar procedure was used to differentiate raters using the formula below.

$$\rho_r^2 = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + \frac{1}{j}(\hat{\sigma}_{res}^2)}$$

Table 1

Estimating Variance Components from a Two-Facet Crossed Design

Source of Variation	Levels	EMS	Variance Component
Teachers (T)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{tr} + k\,\hat{\sigma}^2_t$	$\hat{\sigma}^2_t = \dfrac{(MS_t - MS_{res})}{k}$
Raters (R)	k	$\hat{\sigma}^2_e + \hat{\sigma}^2_{tr} + j\,\hat{\sigma}^2_r$	$\hat{\sigma}^2_r = \dfrac{(MS_r - MS_{res})}{j}$
Residual		$\hat{\sigma}^2_e + \hat{\sigma}^2_{tr}$	$\hat{\sigma}^2_{res} = MS_{res}$

TABLE 2

Analysis of Variance for a
Two Facet Crossed Design

Source of Variation	df	MS	Levels	Variance Component
Teachers (T)	3	1856.92	4	301.89
Raters (R)	5	101.42	6	13.96
Residual	15	45.58		45.58

A generalizability coefficient of .55 was obtained. This is certainly lower than the value of .80 that is often used as a comparison value (Cardinet et al, 1976). The coefficient indicates that the measures are not dependable for differentiating raters. Variation between raters may be too high to be acceptable in many contexts. Thus, the coefficients may provide an indication that an intended set of procedures should be modified in an attempt to increase interrater agreement to lower the generalizability coefficient further.

Nested Design

The design of a generalizability study will determine the coefficients that can be calculated. As a general rule, more information can be obtained from a fully crossed design as coefficients can be estimated for all facets. To introduce the concept of nesting, consider the following example which is a modification of the first design. Six teachers are each assessed by three raters, each rater assessing only one teacher. Thus, each teacher is assessed by a different set of raters. Formulae to estimate variance components for this example are contained in Table 3 and ANOVA data are provided in Table 4.

(Include Tables 3 and 4 About Here)

Since each rater assesses only one teacher, the variance attributable to differences between raters is confounded with variance due to rater by teacher interaction and error variance. As a consequence, a generalizability coefficient to differentiate raters is not estimable.

The generalizability coefficient to differentiate among teachers is provided below. Since decisions on teacher performance will be based on the mean from three raters, the residual error term is reduced by a factor of three.

TABLE 3

Estimating Variance Components from
a Two-Facet Nested Design

Source of Variation	Levels	EMS			Variance Component
Teachers (T)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{r(t)} + k\hat{\sigma}^2_t$	$\hat{\sigma}^2_t$	=	$\frac{(MS_t - MS_{res})}{k}$
Residual	k	$\hat{\sigma}^2_e + \hat{\sigma}^2_{r(t)}$	$\hat{\sigma}^2_{res}$	=	MS_{res}

TABLE 4

Analysis of Variance for a Two-Facet Nested Design

Source of Variation	df	MS	Levels	Variance Component
Teachers (T)	5	596.13	6	193.84
Residual	12	14.61		14.61

$$\rho_t^2 + \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \frac{1}{k}(\hat{\sigma}^2_{RESIDUAL})}$$

Rho-squared in this case was .98. Mean performance across three raters was quite dependable for distinguishing teachers.

THREE FACET DESIGNS

When the competencies are regarded as a source of variance, a third facet is introduced into the design. The procedures used to calculate coefficients in the first two examples are readily extended to incorporate a third facet. In this case, teachers and raters are considered to be random samples from a universe of teachers and raters respectively. In terms of the analysis of variance model, these facets are regarded as random effects. This is not the case with the competencies facet. The competencies define the universe, for they are not considered to be a random sample drawn from an infinite number of similar competencies. This facet is a fixed effect.

The inclusion of a fixed facet in the design can affect the magnitude of generalizability coefficients. Interaction terms for the face of differentiation with a fixed effect are included in the numerator of the equation for estimating the generalizability coefficient (Cardinet et al, 1976). The only implications of a fixed effect are for generalization. As far as differentiating a facet is concerned it makes no difference whether it is random or fixed. According to Cardinet et al:

> Fixing or not fixing the conditions of the facet of generalization has important consequences on the value of the generalizability coefficients. We have not found similar consequences for defining one way or the other the conditions of the facets of differentiation.

A number of the possible generalizability coefficients are discussed in the following sections for the fully crossed design in which six raters assess each of the four teachers on the ten competencies. Table 5 contains

the formulae used to estimate variance components for this design. Table 6 contains variance component estimates applicable to the three facet design specified.

(Include Table 5 and 6 About Here)

Crossed Design

Case 1: Differentiating Teachers, Generalizing over Raters and Competencies. As was the case with the earlier examples, mean values of measures are the sources of differentiation. As a consequence, variance estimates are based on the number of cases used to determine the relevant mean.

$$\rho_t^2 = \frac{\hat{\sigma}_t^2 + \frac{1}{i}\hat{\sigma}_{tc}^2}{\hat{\sigma}_t^2 + \frac{1}{i}\hat{\sigma}_{tc}^2 + \frac{1}{k}\hat{\sigma}_{rt}^2 + \frac{1}{ik}\hat{\sigma}_{res}^2}$$

In this case the generalizability coefficient is .94. Thus teacher performance can be differentiated on the competencies by any of the raters, i.e., the coefficient provides a measure of ability to differentiate teachers and generalize to raters and competencies. It should be noted that interactions of the facet of differentiation with a fixed effect are included in the numerator. The denominator contains all components from the numerator and interaction terms for the facet of differentiation with the facet of generalization.

Case 2: Differentiating Raters, Generalizing over Teachers and Competencies. A similar formula provides an estimate of the ability to differentiate raters and generalize over teachers and competencies.

$$\rho_r^2 = \frac{\hat{\sigma}_r^2 + \frac{1}{i}\hat{\sigma}_{rc}^2}{\hat{\sigma}_r^2 + \frac{1}{i}\hat{\sigma}_{rc}^2 + \frac{1}{j}\hat{\sigma}_{rt}^2 + \frac{1}{ij}\hat{\sigma}_{res}^2}$$

TABLE 5

Expected Mean Squares and Variance Components for a Three-Facet Fully Crossed Design

Source of Variation	Levels	EMS	Variance Component
Teachers (T)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + i\hat{\sigma}^2_{tr} + k\hat{\sigma}^2_{tc} + ki\hat{\sigma}^2_t$	$\hat{\sigma}^2_t = \frac{(MS_t - MS_{tr} - MS_{tc} + MS_{res})}{ki}$
Raters (R)	k	$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + j\hat{\sigma}^2_{rc} + i\hat{\sigma}^2_{tr} + ij\hat{\sigma}^2_r$	$\hat{\sigma}^2_r = \frac{(MS_r - MS_{rc} - MS_{tr} + MS_{res})}{ij}$
Competencies (C)	i	$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + j\hat{\sigma}^2_{rc} + k\hat{\sigma}^2_{tc} + jk\hat{\sigma}^2_c$	$\hat{\sigma}^2_c = \frac{(MS_c - MS_{rc} - MS_{tc} + MS_{res})}{jk}$
T x R		$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + i\hat{\sigma}^2_{tr}$	$\hat{\sigma}^2_{tr} = \frac{(MS_{tr} - MS_{res})}{i}$
T x C		$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + k\hat{\sigma}^2_{tc}$	$\hat{\sigma}^2_{tc} = \frac{(MS_{tc} - MS_{res})}{k}$
R x C		$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc} + j\hat{\sigma}^2_{rc}$	$\hat{\sigma}^2_{rc} = \frac{(MS_{rc} - MS_{res})}{j}$
Residual		$\hat{\sigma}^2_e + \hat{\sigma}^2_{trc}$	$\hat{\sigma}^2_{res} = MS_{res}$

TABLE 6

Analysis of Variance for a
Three Facet Crossed Design

Source of Variation	df	MS	Levels	Variance Component
Teacher	3	137.3087	j = 4	2.0054
Rater	5	9.8576	k = 6	0.0146
Competency	9	277.0730	i = 10	11.1120
Teacher x Rater	15	8.6951		0.7180
Teacher x Competency	27	9.8053		1.3817
Rater x Competency	45	2.0940		0.1448
Residual	131	1.5150		1.5150

In this case a coefficient of .06 was obtained. This clearly provides an indication of low variation in rater assessments, or high interrater agreement.

Case 3: Differentiate Competencies and Generalize Over Teachers and Raters. If the question in an investigation was whether differences existed between competencies for all teachers and raters, the appropriate generalizability coefficient is calculated using the formula provided below.

$$\rho_c^2 = \frac{\hat{\sigma}_c^2}{\hat{\sigma}_c^2 + \frac{1}{k}\hat{\sigma}_{rc}^2 + \frac{1}{j}\hat{\sigma}_{tc}^2 + \frac{1}{jk}\hat{\sigma}_{res}^2}$$

The value obtained for rho-squared in this example was .96. Quite clearly the measures obtained could be used to differentiate among competencies and generalize to all raters and teachers.

Case 4: Differentiate Teachers and Competencies and Generalize over Raters. A generalizability coefficient that provides an estimate of the dependability of performance measures for differentiating teachers and categories for all raters is important as it relates to the way that the TPAI will be used in assessing teacher performance. The formula for calculating rho-squared for this application is provided below.

$$\rho_{t \text{ and } c}^2 = \frac{\hat{\sigma}_t^2 + \hat{\sigma}_c^2 + \hat{\sigma}_{tc}^2}{\hat{\sigma}_t^2 + \hat{\sigma}_c^2 + \hat{\sigma}_{tc}^2 + \frac{1}{k}(\hat{\sigma}_{tr}^2 + \hat{\sigma}_{cr}^2 + \hat{\sigma}_{res}^2)}$$

The value obtained for rho-squared in this example was .97. This indicates that the performance measures can be used as a basis for differentiating teachers and competency level. Measurement of teacher performance is reliable at the competency level.

Case 5: Differentiating Raters and Competencies and Generalizing Over Teachers. There are instances when raters would want to be differentiated in an investigation (e.g., the effects of variations in training on rater effectiveness). The formula for this generalizability coefficient is provided below.

$$\rho^2_{r \text{ and } c} = \frac{\hat{\sigma}^2_r + \hat{\sigma}^2_c + \hat{\sigma}^2_{rc}}{\hat{\sigma}^2_r + \hat{\sigma}^2_c + \hat{\sigma}^2_{rc} + \frac{1}{j}(\hat{\sigma}^2_{rt} + \hat{\sigma}^2_{ct} + \hat{\sigma}^2_{res})}$$

A coefficient of .93 provides an indication that mean teacher performance measures could be used to differentiate raters on specific competencies.

A high coefficient in this case requires further investigation. An investigator may not want high variation between raters or high variation due to a rater-competency interaction. In studies or evaluations where teachers are being differentiated, variance components attributable to these effects should be small.

It has been previously shown (Case 2), that the generalizability coefficient for distinguishing raters was small ($\rho^2 = .06$). The rho-squared calculation for the rater-competency interaction is discussed in the next section.

Case 6: Variation Due to Interaction between Facets. The three facet design contains four interaction terms. The three-way interaction is confounded with random error and cannot be estimated in this design. A teacher by competency interaction is possible and would be expected in the example being considered. In contrast, a rater by teacher interaction, or a rater by competency interaction is also possible but should be small and attributable to random error. Where the interaction is considered "unlawful", the generalizability coefficient should be low. A high coefficient would provide

an indication of change being required in the procedures that are to be adopted.

Two generalizability coefficients can be calculated for the example being considered. It is not possible to calculate the rater by teacher coefficient because of confounding of the error term with the three-way interaction. The data in Table 7 indicate the lawful nature of the teacher x competency interaction and the random nature of the rater x competency interaction.

(Include Table 7 About Here)

Discussion. The models developed above are generalizable to other three facet crossed designs. For example, categories of engagement could be assessed for teachers on multiple occasions. In this case, categories of engagement would represent a fixed facet and teachers and occasions would be regarded as random facets. Similarly, the model is readily extended to incorporate additional facets such as time of day, subject matter, grade level, etc.

Nested Design

As was the case with the two facet design, raters could have been nested within teachers. For example, if six teachers are each assessed by three raters on ten competencies, a nested design is appropriate for estimating variance components. The analysis of variance formulation for this model is contained in Table 8.

(Include Table 8 About Here)

Because there is only one observation per cell, variance attributable to the interaction between competencies and raters nested within teachers is confounded with error variance. In addition, since each rater observes

TABLE 7

Generalizability Coefficients for Variation due to Interactions among Facets

Source of Variation	Formula	ρ^2
Teacher x Competency	$\rho^2_{tc} = \dfrac{\hat{\sigma}^2_{tc}}{\hat{\sigma}^2_{tc} + \frac{1}{k}\hat{\sigma}^2_{res}}$	.85
Rater x Competency	$\rho^2_{rc} = \dfrac{\hat{\sigma}^2_{rc}}{\hat{\sigma}^2_{rc} + \frac{1}{j}\hat{\sigma}^2_{res}}$	.28
Rater x Teacher	Not Calculable with this design	--

TABLE 8

Components of Variance for a Three Facet Nested Design

Source of Variation	Levels	EMS	Variance Components
Teachers (T)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{cr(t)} + i\hat{\sigma}^2_{r(t)} + k\hat{\sigma}^2_{ct} + ik\hat{\sigma}^2_t$	$\hat{\sigma}^2_t = \frac{(MS_t - MS_{r(t)} - MS_{ct} + MS_{res})}{ik}$
Competencies (C)	i	$\hat{\sigma}^2_e + \hat{\sigma}^2_{cr(t)} + k\hat{\sigma}^2_{ct} + jk\hat{\sigma}^2_c$	$\hat{\sigma}^2_c = \frac{(MS_c - MS_{ct})}{jk}$
Raters (Teachers) R(T)	k	$\hat{\sigma}^2_e + \hat{\sigma}^2_{cr(t)} + i\hat{\sigma}^2_{r(t)}$	$\hat{\sigma}^2_{r(t)} = \frac{(MS_{r(t)} - MS_{res})}{i}$
C x T		$\hat{\sigma}^2_e + \hat{\sigma}^2_{cr(t)} + k\hat{\sigma}^2_{ct}$	$\hat{\sigma}^2_{ct} = \frac{(MS_{ct} - MS_{res})}{k}$
Residual		$\hat{\sigma}^2_e + \hat{\sigma}^2_{cr(t)}$	$\hat{\sigma}^2_{res} = MS_{res}$

a teacher only once, there is no way to estimate the variance due to raters independently of the teacher by rater interaction. As a consequence, a generalizability coefficient cannot be calculated for this design to differentiate raters. Analysis of variance data for the above example are contained in Table 9.

As in the crossed three facet design, a number of generalizability coefficients can be calculated. Each of these is calculated and discussed below.

Case 7: Differentiate Teachers and Generalize over Raters and Competencies. The formula to calculate the rho-squared coefficient contains a teacher by competency term in the numerator because the competencies facet is a fixed effect.

$$\rho_t^2 = \frac{\hat{\sigma}_t^2 + \frac{1}{i}\hat{\sigma}_{tc}^2}{\hat{\sigma}_t^2 + \frac{1}{i}\hat{\sigma}_{tc}^2 + \frac{1}{k}\hat{\sigma}_{r(t)}^2 + \frac{1}{ik}\hat{\sigma}_{res}^2}$$

(Include Table 9 About Here)

Rho-squared in this example was .98. Thus, when a nested design such as this is used, the performance measures obtained may be reliably used to differentiate teachers.

Case 8: Differentiate Competencies and Generalize over Teachers and Raters. With this design a rho-squared coefficient can be calculated to differentiate competencies and generalize over teachers and raters. The appropriate formula in this case is listed below.

$$\rho_c^2 = \frac{\hat{\sigma}_c^2}{\hat{\sigma}_c^2 + \frac{1}{j}\hat{\sigma}_{tc}^2 + \frac{1}{jk}\hat{\sigma}_{res}^2}$$

TABLE 9

Analysis of Variance for a
Three Facet Nested Design

Source of Variation	df	MS	Levels	Variance Component
Teacher	5	57.7465	j = 6	1.7378
Rater (Teacher)	12	1.5638	k = 3	0.0732
Competency	9	216.0123	i = 10	11.7296
Teacher x Competency	45	4.8799		1.3492
Residual	97	0.8323		0.8323

The value for rho-squared in this example was .98. This provides an indication of the dependability of the measures for differentiating competency performance averaged over raters and teachers.

Case 9: Differentiating Teachers and Competencies and Generalizing over Raters. Many applications of teacher assessment data require that teachers and competencies are facets of differentiation. A generalizability coefficient which is analogous to that applicable to case 4 can be calculated for a nested design.

$$\rho^2_{t \text{ and } c} = \frac{\hat{\sigma}^2_t + \hat{\sigma}^2_c + \hat{\sigma}^2_{tc}}{\hat{\sigma}^2_t + \hat{\sigma}^2_c + \hat{\sigma}^2_{tc} + \frac{1}{j}\hat{\sigma}^2_{r(t)} + \frac{1}{jk}\hat{\sigma}^2_{res}}$$

The magnitude for rho-squared in this example was .98. Clearly, the data are sufficiently reliable to be used to answer questions requiring differentiation of teachers and competencies.

Case 10: Differentiate the Teacher by Competency Interaction and Generalize across Raters. Logical analysis suggests that a teacher by competency interaction would be a lawful source of variation. If this is the case, a high rho-squared coefficient can be expected. The formula used to calculate the coefficient in this case is shown below.

$$\rho^2_{tc} = \frac{\hat{\sigma}^2_{tc}}{\hat{\sigma}^2_{tc} + \frac{1}{k}\hat{\sigma}^2_{res}}$$

The notion that the teacher by competency interaction represents a lawful source of variation is supported by a high rho-squared value of .83.

DISCUSSION

The notion of reliability being associated with the uses to be made of data is powerful. Measures may be reliable for one set of purposes and quite unreliable for another. As a consequence, generalizability coefficients should be estimated for all questions that have been asked in an investigation. Examples of generalizability coefficients that are applicable to a range of questions are provided in Figure 1. Information on the reliability of the measures for answering specific questions provides an indication on the probability of committing type II errors. If unreliable procedures are used there is a greater probability of failing to reject a false null hypothesis.

(Include Figure 1 About Here)

G-Studies. Exploratory G-studies are advocated to enable generalizability coefficients to be estimated so that procedures can be modified, as necessary, to effect an increase in the dependability of measures obtained for specific uses. For example, assessment of the cognitive level of teacher questioning requires questions to be randomly sampled from a number of lessons. A decision on the number of questions sampled from each lesson has implications for the dependability of the data (see Rowley, 1978) obtained. If teachers (t) are to be differentiated on the basis of mean cognitive level of questions, the formula for the generalizability coefficient to differentiate teachers and generalize over questions (Q) and occasions (O) is provided below.

$$\rho_t^2 = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \frac{1}{O}\hat{\sigma}_{ot}^2 + \frac{1}{Q}\hat{\sigma}_{Q(t)}^2 + \frac{1}{OQ}\hat{\sigma}_{res}^2}$$

Coefficient	Question
ρ_t^2 (Cases 1, 7)	Do teachers in urban schools perform at a higher level than those in rural schools? Is there a relationship between teacher gender and teacher performance?

These questions require teacher performance to be differentiated. Mean scores, computed for each teacher across raters and competencies, could be used as a basis for comparison.

ρ_c^2 (Cases 3, 8)	Is performance on competency four higher than performance on competency five? How is performance on competency four related to performance on competency five?

In this case competencies need to be differentiated. The criteria for comparison are mean scores for each competency computed across teachers and raters.

ρ_r^2 (Case 2)	Are performance ratings from trained raters different from untrained raters? How are performance data from male raters related to those from females?

These questions require that performance ratings given by raters be differentiated. The basis for comparison in this case would be a mean score for each rater computed across teachers and competencies.

$\rho_{t \text{ and } c}^2$ (Cases 4, 9)	Do teachers in urban schools perform at higher levels than teachers in rural schools on classroom procedures competencies? How is performance of urban teachers related to performance of rural teacher on competency 5?

This generalizability coefficient is applicable when teachers are compared at the competency level. The criteria for such comparisons are means for each teacher on each competency across raters.

ρ_{rt}^2	Is there a rater bias toward some teachers?

This coefficient is particularly useful in a G study. A high coefficient may indicate a need for additional training for raters.

Figure 1. Examples of uses of selected generalizability coefficients.

Thus, the effects of changing the number of questions (Q) and/or number of occasions (O) on the magnitude of rho-squared can be estimated from the formula. It is assumed that changes in O and/or Q will not affect the magnitude of the variance components. G-study data can be used in an analogous manner to make decisions on the number of levels for any facet.

The use of a comparatively small, fully crossed G-study can provide estimates that are applicable to larger scale D-studies. If nesting is unavoidable in a G-study (e.g., if categories of student engagement are measured within intact classes), as many facets as possible should be crossed in the study design. The use of a nested design places limits on the number of generalizability coefficients that can be estimated. Even if a D-study is to incorporate a nested design, there are advantages in designing a fully crossed G-study. These advantages relate to the information obtained on the reliability of potential sources of variance. By estimating the magnitude of generalizability coefficients, information is obtained on whether a source of variation is attributable to error or to systematic change. For example, in many situations systematic variation attributable to raters, or a rater by teacher interaction is undesirable. In such cases, estimates of the magnitudes of coefficients to differentiate raters and to differentiate the rater by teacher interaction can be used as a basis for modifying data collection procedures. An indication of a need for change can also be inferred from relatively large negative variance components which indicate a sampling bias.

Nested Designs. Examples of nested designs are numerous in classroom contexts. For example, when student data are collected the design will almost certainly be nested since students are usually nested within

classes. If student engagement is rated on a number of fixed categories by randomly selected raters, a four facet model is applicable. The facets are shown with formulae to estimate variance components in Table 10.

(Include Table 10 About Here)

It should be noted that a generalizability coefficient cannot be estimated for differentiating students because of confounding of variance due to students with variance due to the student by class interaction. Nevertheless, questions about engagement rates of individuals may be a concern. Estimates on the dependability of the measures obtained are therefore desirable. Since a fully crossed G-study may not be practical in such situations, generalizability coefficients to differentiate students can be calculated for each class separately and the range and mean of these values reported.

The actual use to be made of measures in answering questions can affect the design used to estimate variance components. For example, if teachers are assigned to treatment groups in a study, treatment can be regarded as a facet in the design. The example presented in case 1 can be modified such that teachers, raters, competencies and treatments are sources of variation. A generalizability coefficient to differentiate teachers would not be calculable in this case as teachers are nested within treatments. A coefficient to differentiate treatments is calculable, however, and is appropriate for answering questions based on treatment group means. Formulae similar to those provided in Table 10 may be used to estimate variance components of interest for this example.

TABLE 10

Variance Component Estimates for a Four Facet Nested Design

Classes C	i	$\hat{\sigma}^2_c = \frac{(MS_c + MS_{crm} + MS_{ms(c)} + MS_{rs(c)} - MS_{cr} - MS_{cm} - MS_{s(c)} - MS_{res})}{jkl}$
Raters R	j	$\hat{\sigma}^2_r = \frac{(MS_r - MS_{cr} - MS_{rm} + MS_{crm})}{ikl}$
Engagement Mode M	k	$\hat{\sigma}^2_m = \frac{(MS_m - MS_{cm} - MS_{rm} + MS_{crm})}{ijl}$
Student (Class) S(C)	l	$\hat{\sigma}^2_{s(c)} = \frac{(MS_{s(c)} + MS_{res} - MS_{ms(c)} - MS_{rs(c)})}{jk}$
C X R		$\hat{\sigma}^2_{cr} = \frac{(MS_{cr} + MS_{res} - MS_{rs(c)} - MS_{crm})}{kl}$
C X M		$\hat{\sigma}^2_{cm} = \frac{(MS_{cm} + MS_{res} - MS_{ms(c)} - MS_{crm})}{jl}$
R X M		$\hat{\sigma}^2_{rm} = \frac{(MS_{rm} - MS_{crm})}{il}$
C X R X M		$\hat{\sigma}^2_{crm} = \frac{(MS_{crm} - MS_{res})}{l}$
MS(C)		$\hat{\sigma}^2_{ms(c)} = \frac{(MS_{ms(c)} - MS_{res})}{j}$
RS(C)		$\hat{\sigma}_{rs(c)} = \frac{(MS_{rs(c)} - MS_{res})}{k}$
Residual		$\hat{\sigma}^2_{res} = MS_{res}$

CONCLUSIONS

Although discussion in this paper has been limited to comparatively simple designs, the model is readily extended to incorporate additional facets and complex nesting arrangements. For a detailed treatment of such designs the reader is referred to Cronbach et al, 1972.

Generalizability theory provides a framework for incorporating multiple sources of variation into the estimation of reliability coefficients of measures to be used in specified contexts. The generalizability formulation is particularly applicable to educational measures which are often affected by a number of identifiable sources of variance. Because of the contextual dependence of generalizability coefficients, care must be taken to identify relevant sources of variance and to accurately specify the statistical model so that unbiased estimates of the variance components can be made.

The importance of reliability coefficients is in planning investigations and interpreting data. As a consequence, the use of G-studies is advocated to estimate generalizability coefficients as a source of information on the suitability of procedures to be used in D-studies. Where possible, generalizability coefficients should also be estimated for each research or evaluation question investigated in a D-study. Because of the relationship between reliability and the probability of committing type II errors, the magnitude of a generalizability coefficient is of importance in interpreting results where hypothesized relationships were not confirmed.

REFERENCES

Capie, W., Johnson, C. E., Anderson, S. J., Ellett, C. D. & Okey, J. R. Teacher Performance Assessment Instruments. Athens: Teacher Assessment Project, University of Georgia, 1979.

Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

McGaw, B., Wardrop, J. L. & Bunda, M. A. Classroom observation schemes: Where are the errors? American Educational Research Journal, 1972, 9, 13-27.

Rowley, G. The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement, 1978, 15, 165-180.